# Analysis of aptamer sequence activity relationships†

**Mark Platt,‡[ab] William Rowe,‡[ab] Joshua Knowles,[ad] Philip J. Day[ac] and Douglas B. Kell*[ab]**

***Received 26th August 2008, Accepted 15th October 2008***
***First published as an Advance Article on the web 12th November 2008***
**DOI: 10.1039/b814892a**

DNA sequences that can bind selectively and specifically to target molecules are known as aptamers. Normally such binding analyses are performed using soluble aptamers. However, there is much to be gained by using an on-chip or microarray format, where a large number of aptameric DNA sequences can be interrogated simultaneously. To calibrate the system, known thrombin binding aptamers (TBAs) have been mutated systematically, producing large populations that allow exploration of key structural aspects of the overall binding motif. The ability to discriminate between background noise and low affinity binding aptamers can be problematic on arrays, and we use the mutated sequences to establish appropriate experimental conditions and their limitations for two commonly used fluorescence-based detection methods. Having optimized experimental conditions, high-density oligonucleotide microarrays were used to explore the entire loop–sequence–functionality relationship creating a detailed model based on over 40 000 analyses, describing key features for quadruplex-forming sequences.

## Introduction

The development of the technique known as SELEX[1,2] or systematic evolution of ligands by exponential enrichment has yielded the ability to raise polynucleotides with high affinity and specificity to target molecules. These nucleotide sequences, known as aptamers, have been developed to bind targets ranging from small molecules to polypeptides and proteins.[1–4] With binding affinities comparable to those of biologically derived antibodies, the anthropogenic nature of aptamers means that they possess much greater flexibility in terms of applications, with uses encompassing diagnostics and therapeutics.[5,6]

[a] *Manchester Interdisciplinary Biocentre, The University of Manchester, 131 Princess Street, Manchester, UK M1 7DN. E-mail: dbk@manchester.ac.uk*
[b] *School of Chemistry, The University of Manchester, Oxford Road, Manchester, UK M13 9PL*
[c] *School of Translational Medicine, The University of Manchester, Oxford Road, Manchester, UK M13 9PT*
[d] *School of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester, UK M13 9PL*
† Electronic supplementary information (ESI) available: Construction of the statistical model and colour version of Fig. 4. See DOI: 10.1039/b814892a
‡ These authors contributed equally to this work.

Aptamers with affinity to the coagulation protein thrombin were among the first raised to a protein target and to date perhaps represent the most comprehensively studied.[4] Sequencing of DNA aptamers derived from the SELEX process against thrombin reveals the dependency of binding on the consensus sequence; $GGtTGGN_{2-5}GGTtGG$.[4,7] Within this sequence mutual hydrogen bonding between the tetrad of guanine repeats leads to the formation of a unimolecular quadruplex structure in the presence of monovalent cations, as evidenced by NMR[7] and crystallographic structural studies[8] (see Fig. 1). This is not only of interest because it highlights the complex interplay between nucleic acid structure and binding affinity, but such structures offer biological significance, as the formation of G-quadruplexes within genomic DNA has been linked with various processes including transcriptional control[9] and telomeric maintenance.[10,11]

The consensus sequence of thrombin-binding aptamers describes the protein–DNA interaction parameters, and thus if G-quadruplex structure is vital to protein binding the quadruplex structural parameters are also inherent to this model. Understanding the relationship between biological sequence and structure can aid in the detection of putative G-quadruplex structures within the genome,[9] while correlating sequence to binding affinity can be useful to many aspects of

## Insight, innovation, integration

Aptamers are ideal for diagnostic and pharmaceutical studies, but gaining knowledge into mechanism and key structural features is essential for novel and diverse future applications. DNA microarrays allow thousands of sequences to be interrogated simultaneously. We have therefore utilized a high density array format to screen key structural features for G quadruplex forming sequences, using the known protein thrombin. This format rapidly yields a vast amount of data allowing a detailed model to be built describing key loop–sequence–functionality relationships. The ability to survey the landscape systematically using aptamers of known sequence makes microarray formats highly suited for studying sequence specific protein binding profiles.

 

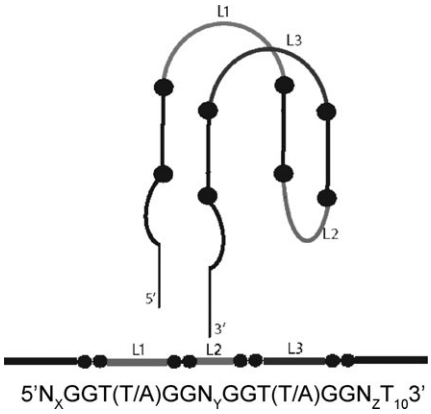


**Fig. 1** Schematic of the G quadruplex structure. Dark circles represent the position of the G bases. Below is the general motif that is typically used to describe TBAs.

aptamer design in a manner analogous to the Quantitative Structure–Activity Relationship (QSAR) assessments used during the design of drug candidates.[12]

SELEX has been used in the characterization of transcription factor binding sites, by the selection and enrichment of DNA sequences with high affinity to the target protein over several iterations. Good binders are sequenced and used to construct a weight matrix to search genomic DNA for potential transcription factor binding sites. However the training of weight matrices using conventional SELEX methodologies is beset with problems, often resulting in very poor correlation between dissociation constants and the weight matrix scores.[13,14] This results from a combination of noise within the SELEX process and over-selection of the best aptamers from each generation.[13] It has been demonstrated that in order to construct an accurate model of DNA binding, information is required not only from good binders but also for those with medium and low affinities.[15] Consequently, many new protocols have been implemented which deliberately select weaker binding sequences.[14,15]

With recent advances in technology both the cost and feature size in DNA microarrays have decreased, and they are now commonly used to perform highly parallel bioanalyses. Through high-density arrayed features, gene expression profiling, comparative genomic hybridization and SNP detection can now be performed quickly and efficiently.[16] High-density microarray technology has advantages over the SELEX technique for characterizing sequence specificity of transcription factor binding sites as sequences can be systematically varied, providing the complete combinatorial landscape of affinity to the protein.[17] The same holds true for understanding the sequence-dependent effectiveness of oligonucleotide hybridization.[18,19] We have applied this generic approach to perform a comprehensive analysis of sequence specificity at each of the loop regions of the thrombin aptamer (see Fig. 1). Using this information and that of over forty thousand variants around the core quadruplex structure, we have built a local structure–activity relationship model which describes the binding affinity of the thrombin aptamer. This also provides insights into the stability of the G-quadruplex structure.

One of the major advantages of the 'on-chip' approach, as highlighted by investigations into the sequence specificity of DNA binding proteins, is that in addition to the removal of the expensive sequencing step, a full range of sequences with differing affinities can be studied.[17] However, this approach purposely includes those variants that have a catastrophic effect on structure and/or binding. In order to characterize these interactions fully it is necessary to optimize the detection method for binding.

The most common detection method used with arrays is fluorescence. The use of fluorochromes however is not without complications, and whilst techniques can label proteins without modifying their underlying structure, there exists a huge body of research showing how the conjugated π-systems possessed by most modern fluorophores used in bioanalysis can interact directly with DNA *via* numerous mechanisms.[20] The direct labelling of the target can therefore enable the dye itself to stabilize or to contribute significantly to the binding, resulting in false positive binding events and high background intensities. We have therefore also attempted to optimize the 'on-chip' detection method in order to develop an accurate model for protein binding.

## Materials and methods

Bovine serum albumin, HABA–avidin reagent, sodium carbonate, sodium bicarbonate, potassium chloride, sodium chloride, Tween 20, sodium dihydrogen phosphate, di-sodium hydrogen phosphate, (>99.5%) and thrombin from human plasma (50–300 NIH units per mg protein) were purchased from Sigma-Aldrich, UK. Cy5 mono reactive dye pack and PD-10 desalting columns were purchased from Amersham Biosciences, UK. 6-((Biotinoyl)amino)hexanoic acid succinimidyl ester and streptavidin-Cy5 (43-4316) were purchased from Invitrogen, UK. Water was obtained from a milli-Q purification system (Millipore, 18 MΩ). Aptamers for SPR analysis were synthesized with a 3′ biotin TEG modification, best scoring aptamer 5′TGGGAGTAGGTTGGTGTGGTTGGGGCTCCCC3′, and ThB sequences were purchased from Eurofins MWG Operon.

Thrombin labelling was carried out according to the standard operating procedure provided from each supplier. Thrombin samples were labelled with either Cy5 dye (th-Cy5) or biotin (th-bio). Purification of the labelled protein from unreacted dye and biotin label was performed on a desalting column equilibrated with hybridization buffer. The ratios of label to protein were calculated using a Nanodrop ND-1000 UV-Vis spectrophotometer. Concentrations of the protein were calculated from known extinction coefficients;[21] it was determined that the average ratio of Cy5 : protein was 3 : 1. The number of biotin moieties per protein was determined using the HABA–avidin test, showing on average a ratio of 3 : 1.

Unless stated all hybridizations were performed in a phosphate buffered saline solution, 1xPBS (0.15 M NaCl, 5 mM KCl, 20 mM phosphate buffer, pH 7.4) at 37 °C. Prior to hybridization all chips were incubated with a prehybridization solution (5% BSA, 0.5% Tween in the hybridization buffer) for 30 minutes, 37 °C. Chips were incubated with the hybridization solutions, th-Cy5 or th-bio, 1xPBS for one hour; the concentration of thrombin in each hybridization was 2 μM; after which the th-Cy5 compartments were washed three times with 1xPBS at room temperature and imaged

immediately. The th-bio compartments had the hybridization solution removed, washed once with 1xPBS before incubating with Strep-Cy5 (1 : 20 dilution from stock) for 2 minutes, washed twice with 1xPBS and imaged immediately.

Fluorescent intensities from the arrays were obtained according to Combimatrix protocols, using a Genepix 4000B scanner (Axon instruments). Image analysis was performed using Combimatrix Microarray Imager software. Data preparation and inspection were carried out using JMP 7.0. Local concentration gradients were removed from chips in R 2.6.0, using the *marray* packages (available from http://www.bioconductor.org), producing scores.

DNA array chips were synthesized in-house on a Combimatrix B3 synthesizer, a detailed synthesis protocol can be found elsewhere.[22] Briefly, a 12 K chip contains 12 544 electrode loci onto which unique sequences are synthesized electrochemically. When used as 4 × 2 K chips each chip in conjunction with the hybridization cap creates four individual compartments, and can perform four unique hybridizations simultaneously. 3584 spots were used for fabrication and quality control purposes leaving 2240 experimental spots per hybridization. 90 K variants of the technology contain 94 928 electrodes, of which 93 310 were used for sequences synthesis and the remaining fulfil fabrication and quality control functions. Individual spots on 12 K formats are separated by a distance of 30 μm, and on the 90 K by 20 μm. Aptamers were present in duplicate within each chip, and each chip was duplicated. All aptamers were synthesized such that the 5′ ends were furthest from the chip surface.

Absolute binding was assayed by Surface Plasmon Resonance (SPR) using a Biacore 3000 machine. Two DNA sequences were immobilized onto a Biacore SA chip (2800 resonance units immobilized); the best mutated sequence detected, and the known original thrombin binder for comparison. Five solutions of thrombin, at concentrations ranging from 1 μM–10 nM, in PBS were passed over the chip at 25 μL $s^{-1}$ at 37 °C: between samples the chip was reconstituted with glycine buffer pH 2.5. The reference cell value was subtracted from all sensograms, and binding values calculated using the BIA evaluation software version 4.1, which was provided with the machine.

## Results and discussion

### Optimizing detection methods

An alternative to direct labelling is to use a biotin label in conjunction with a post-hybridization labelling step with a streptavidin–fluorophore conjugate. The advantage of this approach is that the fluorophore is introduced to the hybridization solution for a significantly shorter period of time, after any interaction between sequence and target has been established. The drawback is that a second wash is required and the introduction of this labelling solution can wash away target molecules from aptamers with weak affinities.

Five thrombin aptamers with known dissociation constants[23,24] were synthesized onto the chips as a method of ensuring coherence between off-chip and on-chip behavior. Variants of two known sequences ('ThA': 5′-AGTCCGT***GGTAGGGCAGGTTGG***GGTGACT3′, ThB: 5′-GGGGAGTA***GGTTGGTGTGGTTGGG***GCTCCCC3′), generated through random mutations (point mutation, insertions and deletions) were used to partially populate two of the compartments within a 12 K feature chip (547 sequences). The remaining spots within these compartments were composed of randomly generated 30mers (565 sequences). The detection of bound thrombin

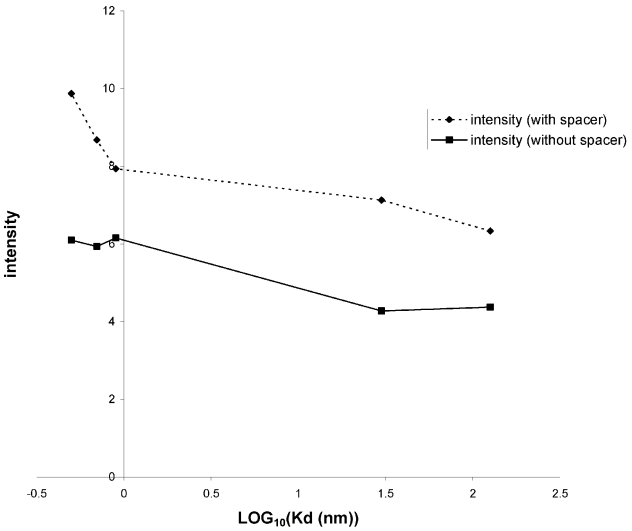


**Fig. 2** Displaying signal intensity on the array of thrombin aptamers against published dissociation constants,[24] with and without T-spacers to balance distance of quadruplex from the chip.

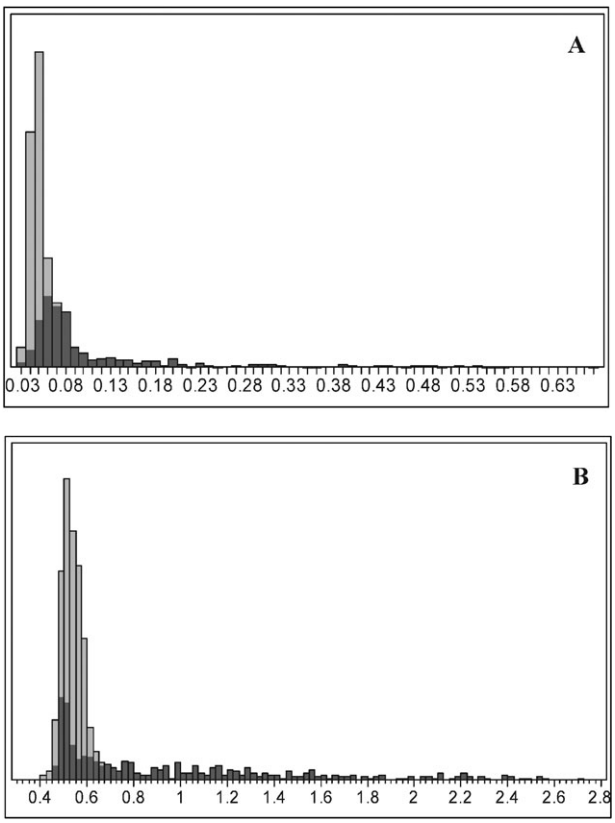


**Fig. 3** Distribution of $\log_{10}$(scores) for both random sequences and sequences containing the motif described in Fig. 1. Highlighted (dark) are the motif scores. (A) Scores obtained using the th-Cy5 method. (B) Scores obtained using the th-bio method.

was then assessed with Cy5 and biotin labels within each of these compartments by comparison of predicted thrombin binding aptamers with the random population. Sequences were predicted to be thrombin binding aptamers if they possessed the consensus sequence $GGtTGGN_{2-5}GGTtGG$ derived from SELEX.

TBAs have been shown to retain their ability to bind thrombin when attached to various surfaces,[25] and this was evident on the Combimatrix B3 platform as all known sequences showed positive binding to both the th-Cy5 and th-bio. As described previously the inclusion of a T-spacer to project the aptamer from the array surface leads to an increase in signal intensity,[26,27] the influence of a poly (T) linker should have minimal effects on structure, but its influence cannot be eliminated totally. As the known TBAs varied in length, to ensure parity the shorter sequences were 'padded out' such that the first quadruplex forming Gs, closest to the 3′ end, were always the same distance from the surface. Binding scores compared to the known dissociation constants are displayed in Fig. 2. Fig. 2 shows that the rank order of binding on and off the array when the T-spacer is present, but not when it is absent. Although based on only five points this equates to a correlation coefficient of 0.9.

While replicating 'off-chip' trends is an important feature, this is meaningless if the detection mechanism is beleaguered with noise from spurious interactions between the label and DNA. As we would expect aptamers with the consensus sequence to have higher affinities to thrombin, these sequences should be discernible from the random population. Fig. 3 displays the distribution of binding scores for sequences containing the consensus sequence; $GGtTGGN_{2-5}GGTtGG$ and the distribution of binding scores of the random population. It is evident that the separation and spread of scores for th-bio are far greater than those corresponding to th-Cy5 (the enhancement of the signal for th-bio is most likely a result of each streptavidin being labelled with more than one dye molecule). Although biotin labelling is able to clearly discriminate strong binders from random, it is difficult to identify those sequences with moderate binding affinities. As these sequences are vital to the generation of a structure–activity relationships model this indicates that biotin–streptavidin may be unsuitable for this task.

The consensus sequence is an indicator of which aptamers are expected to have high binding affinities; however, it is not inherent to all strong binding variants on the chip. The aptamers listed in Table 1 reveal that there is far greater sequence plasticity in loops $L_1$ and $L_3$ than those derived through SELEX. These findings complement previous studies into the stability of quadruplexes where it was seen that the "core" GG repeat unit is essential[25,28] whilst changes in all three loop regions can be made with varying degrees of success. Any mutations in this repeat set "GG" usually cause catastrophic effects, for example a single deletion at point 12 in ThA, drops the rank of the starting sequence from 46th to 324th.

One sequence that scores higher than any other aptamer and that appears in both detection methods is: TGGGAG-TAGGTTGGTGTGGTTGGGGCTCCCC. This sequence

**Table 1** Alignments of sequences with top 5% of scorers which vary from the GGT(T/A)GGN3GGT(T/A)GG motif (top 10% are listed in Table 1 of ESI†)

| Sequence 5′–3′ | Score[a] | Percentile |
|---|---|---|
| TGGGAGTA**GG**TGT**GG** TGT**GG**T T**GG**GGCTCCCC | 76.27 | 1 |
| GTGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC | 66.56 | 2 |
| GGGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GTCTCCCC | 40.79 | 2 |
| GGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GGCTCCCC | 29.21 | 5 |
| GGTGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC | 26.69 | 5 |
| GGGAGTA**GG**T T**GG**TTGT**GG**T T**GG**GGCTCCTC | 21.82 | 5 |
| GGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GGCTCCCC | 14.33 | 5 |
| GGGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GCTCCC | 13.18 | 5 |
| GGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCC | 11.55 | 5 |
| GCGGAGTA**GG**T T**GG**TCGT**GG**T T**GG**GGCTCCCA | 11.32 | 5 |
| | Score[b] | Percentile |
| GTGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC | 1.91 | 2 |
| GGTGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCCC | 1.89 | 5 |
| GGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GGCTCCCC | 1.75 | 5 |
| GGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GGCTCCCC | 1.72 | 5 |
| TGGGAGTA**GG**TGT**GG** TGT**GG**T T**GG**GGCTCCCC | 1.63 | 5 |
| GGGAGTA**GG**T T**GG** GT**GG**T T**GG**GGCTCCC | 1.47 | 5 |
| GGGGAGTA**GG**T T**GG** TGT**GG**TTT**GG**GCTCCCC | 1.41 | 5 |
| GGGAGTA**GG**T T**GG**TTGT**GG**T T**GG**GGCTCCTC | 1.40 | 5 |
| GGGGAGTA**GG**TTT**GG** TGT**GG**T T**GG**GTCTCCCC | 1.39 | 5 |

[a] Scores th-Bio. [b] Scores th-cy5.

arises from mutations to the duplex region of the aptamer rather than the core quadruplex. It is important to highlight that some mutations here may be beneficial to surface analysis and experimental conditions, but not necessarily increase binding constant and inhibition rates. Binding constants calculated from surface plasmon resonance experiments, SPR, are evidence of this fact as both the mutated and original aptamer (ThB) have $K_D$'s of 28 and 26 nM (data shown within ESI†). A previous study which explored the sequence space of immunoglobulin E aptamer using high density microarrays also discovered improved aptamers by directed mutations to a sequence derived from SELEX.[27] In this instance the authors demonstrated that the effect was due to destabilization of the stem region within the single loop aptamer. This improved binding affinity was replicated 'off-chip' using surface plasmon resonance indicating the validity of the chip based approach.

The ability to measure and detect the interaction between protein and aptamer sequence is strongly influenced by the detection mechanism. The development of this technique to allow the study of protein targets that are either novel or poorly studied will rely on the ability to distinguish signal from background noise. However the use of array based formats offer an experimental technique capable of generating a wealth of knowledge.

### Combinatorial landscape

In order to probe any enhancement to binding that can arise from changes in the loop sequences and lengths, all possible combinations of sequences at loops $L_1$ and $L_3$ with lengths of two and three bases were systematically screened on a 90 k chip. These variants were derived from thrombin aptamer ThB ($GGGGAGTAGG(X_{2-3})GGTGTTGG(X_{2-3})GGGGCTCCCC$ where X denotes the bases varied). 41000 pseudo-random variants of the complementary ends and loop $L_2$ around the central quadruplex structure $X_8GGTTX_{2-4}GGTTGGGGX_6$ (where X denotes the bases varied) were also generated. Rather than generating the ends completely randomly there was an 85% probability that each randomly generated base on the 5′ end was paired with a complementary base at the 3′ end; this is because it has been previously shown that stability of duplex strand aids quadruplex stability.[7] In addition loops $L_1$ and $L_3$ were simultaneously varied with 0.01% probability that each loop was extended to a length of three and a 0.001% probability that the each base varied from being a T.

Lengths of loops and complementary sections were observed to have a relatively small effect on binding scores, as too did the sequence composition of the complementary regions. The sequences within each loop, however, showed huge variation in scores (see Fig. 4), with the most important features being prevalence of Gs within loop 1, the presence of "TA" within loop 3, and a sequence with TAG within loop 2. It is important to note that whilst we fix the initial repeat "Gs", substitution of neighboring bases to guanine residues could cause a shifting of the start point of the loop. The increase in Gs in loop 1 may therefore correspond to a stabilization of the quadruplex structure rather than their being present in the loop region.

### Structure–activity relationship model

Conventionally weight matrices derived from SELEX experiments express the prevalence of bases at each position in terms of information content. Using the array platform and obtaining information on a wider range of binding sequences a variety of statistical models can be derived from the data. Linear regression models are limited in that they cannot detect correlations between features within the dataset *i.e.* connectivity between structural features within the TBAs. However unlike many machine learning methods the output is easy to visualize. A linear regression model was generated using the Akaike criterion for model selection,[29] and the model was

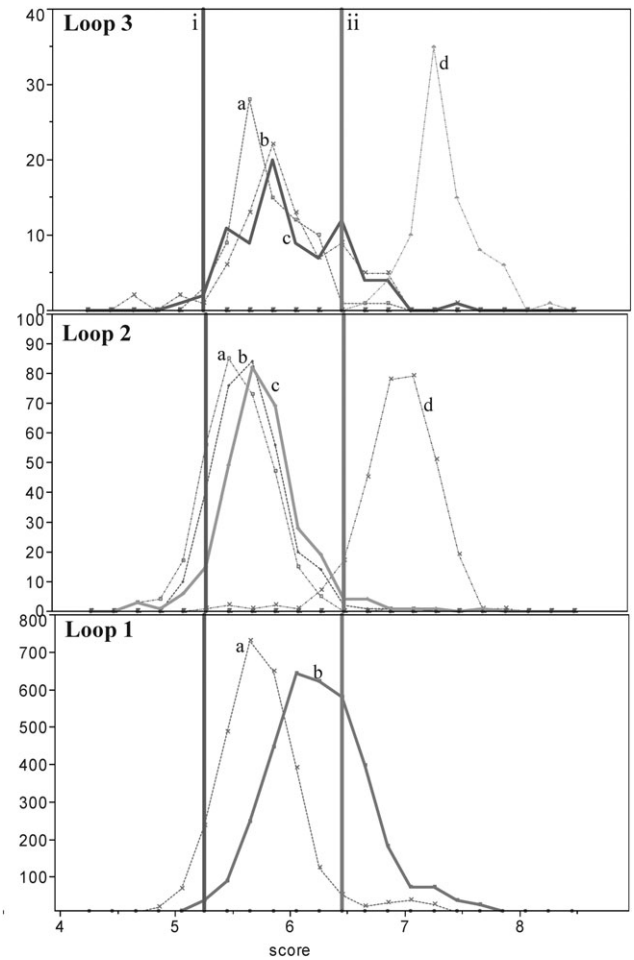


**Fig. 4** Relative scores of sequences that represent key aspects detected within the model; solid vertical line i represents the average score of known thrombin binders[29] with $K_d$'s 30, 42, and 126 nm; vertical line ii, average scores for known thrombin binders with $K_d$'s 0.5, 0.7, 0.9 nm. Loop 1: a—dashed line, sequences that do not contain a guanine residues; b—solid line represents sequences that contain at least one G within loop 1. Sequence for loop 2 remains fixed. Loop 2: a—dot-dashed line represents loop 2 sequence TAT; b—dashed line, loop 2 sequence of TTG; c—solid line, loop 2 sequence CAG; d—dot-dot-dashed line, loop 2 sequence TAG. Loops 1 and 3 remain fixed at TT. Loop 3: a—dashed line, loop 3 sequence GG; b—dot-dashed line, loop 3 sequence AA; c—solid line, loop 3 sequence TT; d—dot-dot-dashed line, loop 3 sequence TA. *N.B.* All permutations of loop sequences were analyzed; shown above is representative subset of data (colour version available in ESI†).

tested using 10-fold cross validation producing a correlation coefficient between observed and predicted scores of 0.71.

It is important to stress that this is a local sequence activity model based around the central quadruplex structure. Predictions from this model may be limited when applied to regions of the sequence space that are well outside the original dataset. While conventionally the thrombin aptamer is constructed around two stacked guanine quartets, it is believed the stability derived from π-orbital interactions from three stacked guanine quartets is desirable in genomic DNA.[30] In fact the thrombin aptamer would not be identified as being a quadruplex using the conventional motif ($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) used for searching for putative structures.[9] Despite this, structural studies of thrombin aptamer have provided insights into the conformational flexibility of G-quadruplexes[7] and serve as a useful model system. Details of the statistical model built on the data in this study are available from the ESI† and raw data are available from http://dbkgroup.org/direvol.htm.

## Conclusions

We show that it is possible to perform on-chip analysis of mutations to the complex three-dimensional thrombin aptamers and derive a statistical model based on these results with a high correlation between predicted and observed binding scores. Whilst the analysis has been performed on a chip, the results reflect previous studies of solution-based assays demonstrating the scope of this technique, with the attendant advantages that arrays permit high-throughput and multiplexing. The chosen method of detection can however limit the quality of the analysis for intermediate binders. Post-hybridization labelling of the target with a streptavidin-fluorophore has been shown to allow simple and ready detection of strong binders with scores and intensities greatly above those of the background and of random sequences. This technique however, requires additional washing of the chip which tends to remove weak binders. Alternatively direct labelling with Cy5 is less cumbersome and allows a more detailed study of intermediate binders. Whilst having the disadvantage that intensities from background and randomly generated sequences can be similar to those of bound targets, analyses indicate that this method is better for differential inspection of mutated sequences.

While the linear regression model described in this paper is able to predict binding affinity based on a feature set used to describe each thrombin aptamer with relatively high accuracy it is unable to detect relationships between features. There exists a plethora of techniques which extend regression to non-linear multivariate models, including neural networks which are rapidly gaining popularity in DNA sequence analysis.[31] These techniques can detect complex relationships between features, which is important when considering structural features, however unlike conventional weight matrices and the linear regression presented here the resultant models are opaque and difficult for human beings to understand. QSAR models for aptameric sequences present an interesting variant in the world of DNA sequence analysis in that unlike the recognition of transcription factor binding sites these are artificial systems. There is no debate about the relationship between *in vitro* and *in vivo* model performance, the accuracy of the model given here should translate when performing rational alternations of the sequence for modification of performance.

The model derived in this study describes the interaction of the protein thrombin with a series of sequences derived from known thrombin aptamers. Central to these sequences is the ability to form G-quadruplex structure. Discerning the features which are critical to quadruplex formation and those which are specific to protein binding is however not an easy task. The DNA microarrays described in this study are capable of being washed and reconstituted; potentially they can be employed again to study the interaction profile of other proteins that bind G-quadruplexes. Similarly two colour dye assays can be employed to study the interaction of two proteins with a single aptamer simultaneously.

As in a previous study on the sequence landscape of an aptamer by randomly varying a sequence produced through SELEX we have observed aptamers on a chip with increased binding affinities.[27] The ability to survey the landscape systematically using aptamers of known sequence makes the 'on-chip' approach considerably more powerful than SELEX for studying sequence specific protein binding profiles. The power of SELEX is in contrast derived from the ability to saturate the sequence landscape to produce high affinity aptamers. The present findings may also indicate that it is best to optimize aptamers derived from SELEX for use in array based applications 'on-chip'.

## Acknowledgements

The authors would like to thank David Broadhurst and Kieran Smallbone for useful discussions and Chris Knight for input on microarray normalization. Joshua Knowles is supported by a David Phillips Fellowship from BBSRC. This work was sponsored by the Biotechnology and Biological Sciences Research Council [PBB/D000203/1], and contributions from the Manchester Centre for Integrative Systems Biology (www.mcisb.org/).

## References

1 A. D. Ellington and J. W. Szostak, *Nature*, 1990, **346**, 818–822.
2 C. Tuerk and L. Gold, *Science*, 1990, **249**, 505–510.
3 F. Ylera, R. Lurz, V. A. Erdmann and J. P. Furste, *Biochem. Biophys. Res. Commun.*, 2002, **290**, 1583–1588.
4 L. C. Bock, L. C. Griffin, J. A. Latham, E. H. Vermaas and J. J. Toole, *Nature*, 1992, **355**, 564–566.
5 S. Tombelli, M. Minunni and M. Mascini, *Biomol. Eng.*, 2007, **24**, 191–200.
6 Y. Li, H. J. Lee and R. M. Corn, *Nucleic Acids Res.*, 2006, **34**, 6416–6424.
7 R. F. Macaya, P. Schultze, F. W. Smith, J. A. Roe and J. Feigon, *Proc. Natl. Acad. Sci. U. S. A.*, 1993, **90**, 3745–3749.
8 K. Padmanabhan, K. P. Padmanabhan, J. D. Ferrara, J. E. Sadler and A. Tulinsky, *J. Biol. Chem.*, 1993, **268**, 17651–17654.
9 J. L. Huppert and S. Balasubramanian, *Nucleic Acids Res.*, 2005, **33**, 2908–2916.
10 R. Giraldo and D. Rhodes, *EMBO J.*, 1994, **13**, 2411–2420.
11 S. D. Patel, M. Isalan, G. Gavory, S. Ladame, Y. Choo and S. Balasubramanian, *Biochemistry*, 2004, **43**, 13452–13458.
12 T. I. Oprea, *Cheminf. Drug Discovery*, Wiley-VCH, 2004.
13 M. Djordjevic, *Biomol. Eng.*, 2007, **24**, 179–189.
14 J. Liu and G. D. Stormo, *Nucleic Acids Res.*, 2005, **33**, e141.

15 E. Roulet, S. Busso, A. A. Camargo, A. J. G. Simpson, N. Mermod and P. Bucher, *Nat. Biotechnol.*, 2002, **20**, 831–835.
16 J. Griffiths, *Anal. Chem.*, 2007, **79**, 8833–8837.
17 C. L. Warren, N. C. S. Kratochvil, K. E. Hauschild, S. Foister, M. L. Brezinski, P. B. Dervan, G. N. Phillips and A. Z. Ansari, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 867–872.
18 N. de-los-Santos-Álvarez, M. a. J. Lobo-Castañón, A. J. Miranda-Ordieres and P. Tuñón-Blanco, *TrAC Trends Anal. Chem.*, 2008, **27**, 437–446.
19 M. Sohail, H. Hochegger, A. Klotzbucher, R. L. Guellec, T. Hunt and E. M. Southern, *Nucleic Acids Res.*, 2001, **29**, 2041–2051.
20 B. A. Armitage, in *DNA Binders and Related Subjects*, ed. M. J. Waring and J. B. Chaires, Springer, Berlin, 2005, vol. 253, pp. 55–76.
21 J. W. Fenton Ii, M. J. Fasco and A. B. Stackrow, *J. Biol. Chem.*, 1977, **252**, 3587–3598.
22 A. L. Ghindilis, M. W. Smith, K. R. Schwarzkopf, K. M. Roth, K. Peyvan, S. B. Munro, M. J. Lodes, A. G. Stover, K. Bernards, K. Dill and A. McSheam, *Biosens. Bioelectron.*, 2007, **22**, 1853–1860.
23 K. Ikebukuro, W. Yoshida, T. Noma and K. Sode, *Biotechnol. Lett.*, 2006, **28**, 1933–1937.
24 D. M. Tasset, M. F. Kubik and W. Steiner, *J. Mol. Biol.*, 1997, **272**, 688–698.
25 A. Bugaut and S. Balasubramanian, *Biochemistry*, 2008, **47**, 689–697.
26 P. J. R. Day, P. S. Flora, J. E. Fox and M. R. Walker, *Biochem. J.*, 1991, **278**, 735–740.
27 E. Katilius, C. Flores and N. W. Woodbury, *Nucleic Acids Res.*, 2007, **35**, 7626–7635.
28 I. Smirnov and R. H. Shafer, *Biochemistry*, 2000, **39**, 1462–1468.
29 H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd edn, Morgan Kaufmann, San Fransisco, USA, 2005.
30 J. E. Johnson, J. S. Smith, M. L. Kozak and F. B. Johnson, *Biochimie*, 2008, **90**, 1250–1263.
31 B. Demeler and G. W. Zhou, *Nucleic Acids Res.*, 1991, **19**, 1593–1599.

 